UNIVERSITE DE MONTPELLIER — FACULTE DES SCIENCES

ANALYSE DES COMPOSANTES PRINCIPALES
UTILISATION DES GROUPES DE VARIABLES
DANS LA RECHERCHE DE LA SOLUTION

T H E S E

présentée à la

FACULTE des SCIENCES de l'UNIVERSITE de MONTPELLIER

pour l'obtention du

DOCTORAT de SPÉCIALITÉ MATHEMATIQUES (3e Cycle)

soutenue le 17 Juin 1966, devant la Commission d'Examen

par

Yves ESCOUFIER

a obtenu Mention Très Honorable)

Jury : Président : Madame **M. LAFON**

Examinateurs : Monsieur **N. ROBY**

Monsieur **J.P.LAFON**

Que Madame le Professeur LAFON veuille bien trouver ici l'expression de ma profonde gratitude pour l'aide qu'elle m'a apportée dans la réalisation de ce travail.

Je remercie Monsieur le Professeur FALGUEIRETTES qui m'a permis d'accéder au laboratoire de calcul de la Faculté aussi souvent qu'il m'a été nécessaire et Monsieur le Professeur EMBERGER qui a autorisé les ingénieurs du C.E.P.E. à me fournir les données que j'ai traitées en exemple.

Je suis très honoré que Messieurs les Professeurs ROBY et LAFON aient accepté de constituer le jury.

Enfin, je tiens à exprimer ma reconnaissance aux personnes du laboratoire de Calcul et à celles du secrétariat de la section de Mathématiques qui se sont chargées des problèmes matériels liés à la réalisation de ce travail.

SOMMAIRE

Introduite au début de ce siècle par Pearson [20] et Hotelling [11], l'analyse des composantes principales est l'une des méthodes d'analyse multivariable les plus employées. Très voisine, par les outils mathématiques qu'elle manipule, de la méthode factorielle proposée par Spearman, elle ne prétend pourtant pas répondre à la même question : si l'analyse factorielle veut reproduire les corrélations existant entre p variables à partir de $m < p$ facteurs, l'analyse des composantes principales veut reproduire les valeurs prises par les variables ([13] P. 33)

En abordant ce travail, je voulais mettre en évidence ce qu'apporte l'hypothèse de normalité des variables. Pour cela, je me suis efforcé, dans le premier chapître, de donner de la méthode une présentation purement géométrique, alors que, en admettant la normalité des variables, je retrouve, dans le second chapître, les formes classiques d'exposition de la méthode et les résultats commandés par cette hypothèse.

Ayant souvent buté sur les difficultés que rencontre l'utilisateur pour interpréter l'analyse, j'introduis, dans le troisième chapître, des résultats simples qui me semblent susceptibles de faciliter l'interprétation.

CHAPITRE PREMIER

I_1- Ayant effectué la mesure de p caractères sur n individus, nous pouvons identifier chaque individu à l'ensemble des p mesures effectuées sur lui. Si nous convenons de repérer les individus par un indice j ($j = 1 \ldots n$) et si x_{ij} est la mesure du caractère i ($i = 1 \ldots p$) sur l'individu j, ce dernier se trouve représenté par l'ensemble $(x_{1j}, x_{2j}, \ldots, x_{pj})$.

Proposons-nous de construire une représentation géométrique des n individus. Dans un espace à p dimensions, défini par un système d'axes orthogonaux, nous identifions chaque axe à l'un des caractères mesurés. Dans cet espace, l'individu j est représenté par le point de coordonnées $x_{1j}, x_{2j}, \ldots, x_{pj}$.

Si p est égal à 1, à 2, ou même à 3, il sera facile de faire cette représentation, dans laquelle la proximité des points sera indicatrice de la ressemblance des individus. Quand le nombre des caractères mesurés augmente, cette méthode n'est plus applicable. Nous pouvons espérer alors que le nuage des points représentant les individus n'est pas isotrope, mais qu'au contraire il s'allonge suivant certaines directions privilégiées alors que sa projection sur d'autres directions est pratiquement nulle. En nous limitant à ces directions privilégiées, nous pourrons construire de notre nuage une représentation approchée qui nous satisfera dans la mesure où l'approximation ne sera pas trop grande. Nous verrons plus loin comment nous pouvons mesurer l'approximation effectuée. (Cf. I_6).

I_2- Introduisons dans l'espace à p dimensions, le produit scalaire habituel et la norme qui en découle. Si X et Y sont deux vecteurs

colonnes $\begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}$ et $\begin{bmatrix} Y_1 \\ \vdots \\ Y_p \end{bmatrix}$, le produit scalaire est défini par :

$$X'Y = X_1Y_1 + \ldots + X_pY_p$$

où X' est le vecteur ligne transposé de X et la norme $\|X\|$ de X est définie par :

$$\|X\|^2 = X_1^2 + \ldots + X_p^2$$

Avec ces notations l'individu j s'identifie à l'extrémité du vecteur $\begin{bmatrix} x_{1j} \\ \vdots \\ x_{pj} \end{bmatrix}$ dont l'origine est à l'origine des axes.

Soit $\overline{x}_i = \dfrac{\sum_{j=1}^{n} x_{ij}}{n}$, la moyenne arithmétique de toutes les mesures faites du caractère i $(i = 1, \ldots, p)$. Faisons subir à l'origine des axes une translation de vecteur $\begin{bmatrix} \overline{x}_1 \\ \vdots \\ \overline{x}_p \end{bmatrix}$. Dans le nouveau système d'axe, l'individu j s'identifie à l'extrémité du vecteur $\begin{bmatrix} x_{1j} - \overline{x}_1 \\ \vdots \\ x_{pj} - \overline{x}_p \end{bmatrix} = \begin{bmatrix} X_{1j} \\ \\ X_{pj} \end{bmatrix}$ que nous conviendrons d'appeler $X_{\cdot j}$.

Soit X la matrice à p lignes et n colonnes obtenue en juxtaposant les n vecteurs colonnes $X_{\cdot j}$.

Le carré de la distance du point représentant l'individu j au point moyen est égal à $\sum_{i=1}^{p} (X_{ij})^2$. La somme des carrés des distances pour les n individus est donc : $D = \sum_{j=1}^{n} \sum_{i=1}^{p} (X_{ij})^2$. La quantité D est liée aux unités choisies pour mesurer les différents caractères, mais ce choix fait, elle est un invariant du système. (Cf I_5)

Donnons nous un autre système orthogonal de référence, de même origine que le précédent. Appelons $V_1, V_2, \ldots, V_p$ les vecteurs unités des différents axes. La coordonnée de l'individu j sur l'axe de vecteur unité V_i est donnée par le produit scalaire

$$V_i' \cdot X_{\cdot j} = Y_{ij} \qquad I_{2.1}$$

Dans ce sytème $D = \sum_{j=1}^{n} \sum_{i=1}^{p} (Y_{ij})^2 = \sum_{i=1}^{p} \sum_{j=1}^{n} (Y_{ij})^2$

Choisir les directions privilégiées dont nous avons parlé plus haut, revient à choisir V_1 parmi les vecteurs unités passant par l'origine de telle sorte que $\sum_{j=1}^{n} (Y_{1j})^2$ soit maximum, puis V_2 parmi les vecteurs unités orthogonaux à V_1 et passant par l'origine de telle sorte que $\sum_{j=1}^{n} (Y_{2j})^2$ soit maximum. En itérant ce procédé, nous construisons un système orthogonal de référence, de même origine que le système initial dans lequel sinon D, du moins un fort pourcentage de cette quantité sera absorbé par $r < p$ axes. L'expression $I_{2.1}$ définit des variables Y_i $(i = 1, \ldots, p)$, prenant la valeur Y_{ij} sur l'individu j. Ce sont les variables Y_i que l'on appelle composantes principales.

I_3- Nous devons choisir V_1 de telle sorte que $\sum_{j=1}^{n} (Y_{1j})^2$ soit maximum. Or : $\sum_{j=1}^{n} (Y_{1j})^2 = \sum_{j=1}^{n} (V'_1 . X_{.j})^2$

et : $I_{3.1}$ $\sum_{j=1}^{n} (V'_1 . X_{.j})^2 = V'_1 . \sum_{j=1}^{n} (X_{.j} X'_{.j}) V_1$.

Le résultat du produit du vecteur colonne $X_{.j}$ par le vecteur ligne $X'_{.j}$ est une matrice A^j d'éléments $a^j_{tu} = X_{tj} \times X_{uj}$. Appelons A la matrice $\sum_{j=1}^{n} (X_{.j} X'_{.j})$ d'éléments $\sum_{j=1}^{n} a^j_{tu}$. A est une matrice à p lignes est p colonnes obtenue en faisant le produit de la matrice X par sa transposée. On sait que A est une matrice symétrique définie positive ([23] p.18).

Pour choisir V_1 de telle sorte que $V'_1 A V_1$ soit maximum avec la condition $V'_1 . V_1 = 1$, nous utiliserons la méthode du multiplicateur de Lagrange, ([22] p. 844), qui nous conduit à écrire que la dérivée par rapport à V_1 de la quantité $2T_1$ doit être nulle où :

$$2T_1 = V'_1 A V_1 - \lambda(V'_1 V_1 - 1)$$

Soit $I_{3.2}$: $\dfrac{\partial T_1}{\partial V_1} = AV_1 - \lambda V_1 = (A - \lambda I)V_1 = 0$

Nous savons que les solutions non triviales de $I_{3.2}$ sont les vecteurs propres de la matrice A.

De $AV_1 - \lambda V_1 = 0$, nous tirons :

$$V_1' A V_1 - \lambda V_1' V_1 = 0$$

Ayant choisi V_1 normé, nous avons donc :

$$I_{3.3} \qquad V_1' A V_1 = \lambda$$

Il ressort donc que nous devons prendre pour V_1 le vecteur propre normé associé à la plus grande valeur propre de A qu'on sait être réelle et positive puisque A est définie positive.

I_4- De manière générale, lorsque nous aurons démontré que les r premières composantes principales sont définies par les r vecteurs propres associés aux r plus grandes valeurs propres, nous devrons choisir V_{r+1} de telle sorte que :

$$V'_{r+1} V_{r+1} = 1$$

$$V'_{r+1} V_i = 0 \qquad i = 1, \ldots, r$$

et nous devrons annuler la dérivée par rapport à V_{r+1} de :

$$2T_{r+1} = V'_{r+1} A V_{r+1} - \lambda(V'_{r+1} V_{r+1} - 1) - 2 \sum_{i=1}^{r} \mu_i V'_{r+1} V_i$$

Soit

$$I_{4.1} \qquad \frac{\partial T_{r+1}}{\partial V_{r+1}} = A V_{r+1} - \lambda V_{r+1} - \sum_{i=1}^{r} \mu_i V_i = 0$$

En multipliant à gauche par V_j' $(j = 1, \ldots, r)$, on obtient

$$V_j' A V_{r+1} - \lambda V_j' V_{r+1} - \sum_{i=1}^{r} \mu_i V_j' V_i = 0$$

Or : $V_j' A = \lambda_j V_j'$ d'où :

$$V_j' V_{r+1} (\lambda_j - \lambda) - \mu_j V_j' V_j = 0$$

Par hypothèse $V_j' V_{r+1} = 0$ et $V_j' V_j = 1$ donc $\mu_j = 0$; si bien que $I_{4.1}$ se réduit à : $(A - \lambda I) V_{r+1} = 0$

On voit alors que l'on doit prendre pour V_{r+1}, le vecteur propre associé à la $(r+1)^{ième}$ valeur propre.

I_5- Que deviennent les composantes principales dans un changement d'unités?

Soit Δ la matrice diagonale qui définit le changement d'unités. Le vecteur $Y_{.j}$ attaché à l'individu j dans le nouveau système est donné par :

$$Y_j = \Delta X_j$$

Nous sommes ramené à chercher des vecteurs normés $W_1, W_2, \ldots, W_p$ qui rendent maximum $\sum_{j=1}^{n} (W_i' Y_j)^2$.

Or : $\sum_{j=1}^{n} (W_i' Y_j)^2 = W_i' \sum_{j=1}^{n} (\Delta X_j X'_j \Delta) W_i = W_i' \Delta (\sum_{j=1}^{n} X_j X_j') \Delta W_i$

Posons $V_i^* = \Delta W_i$. On a immédiatement $\Delta^{-1} V^*_i = W_i$ et donc $V_i^{*\prime} \Delta^{-2} V_i^* = 1$.

Nous sommes donc ramenés à chercher les vecteurs V_i^* qui rendent maximum la quantité $V_i^{*\prime} A V_i^*$ avec pour condition $V_i^{*\prime} \Delta^{-2} V_i^* = 1$

La condition de Lagrange nous donne :

$$(A - \lambda^* \Delta^{-2}) V_i^* = 0$$

On voit que les V_i^* ne sont reliés de façon simple aux V_i que si Δ^{-2} est un multiple de la matrice identité. DAns les autres cas, il est difficile de prévoir les conséquences d'un changement d'unités : c'est là la faiblesse de la méthode des composantes principales.

I_6- Remarquons que la trace de la matrice A n'est rien d'autre que la quantité $D = \sum_{i=1}^{p} \sum_{j=1}^{n} (X_{ij})^2$.

La trace de la matrice étant invariante dans la diagonalisation, on peut en avoir la valeur en faisant la somme des valeurs propres ([25] p.225).

Cette propriété donne une mesure de l'approximation effectuée lorsqu'on remplace l'ensemble des p variables observées par r composantes

principales. Soit Σ_r la somme des r premières valeurs propres. $\frac{\Sigma_r}{D}$ est le pourcentage de la dispersion absorbée par les premières composantes. Si ce pourcentage est grand, la représentation est satisfaisante.

I$_7$- a) Les données que nous traitons proviennent de relevés effectués en Sologne par Monsieur Grandjouan du Centre d'étude Phytosociologique et écologique de Montpellier. Nous le remercions d'avoir bien voulu les mettre à notre disposition.

Nous avons retenu 6 variables (Humidité de la terre p. F 2,7 ; p F 3 ; p F 4,2 ; Epaisseur de la colonne de terre ; N total ; carbone organique) qui donne la matrice $\Lambda^* = \frac{A}{n-1}$ suivante (dont je n'écris que la partie supérieure par commodité).

129 972,80	106 292,57	52 146,90	1 440,29	7197,07	99 669,65
	90 474,66	44 896,89	1 223,98	6346,38	89 719,74
		26 286,94	642,58	3617,71	52 967,02
			27,31	95,53	1 466,86
				899,34	13 897,24
					248138,58

$= \Lambda^*$

La quantité D est ici égale à : 495 799,63.

Le calcul effectué par la méthode de Jacobi (chapitre IV) donne les valeurs propres suivantes :

L(1) =	389 050,33
L(2) =	101 086,27
L(3) =	3 956,41
L(4) =	1 607,28
L(5) =	89,51
L(6) =	10,08
Total	495 799,88

La trace s'est donc conservée aux erreurs d'arrondi près.

La première valeur propre absorbe 79 % de la variation.

Les deux premières en absorbent 99 %. Une représentation sur les deux composantes principales correspondantes est donc très satisfaisante.

Ces composantes principales sont :

$$V_1 = 0{,}50\ X_1 + 0{,}43\ X_2 + 0{,}23\ X_3 + 0{,}01\ X_4 + 0{,}04\ X_5 + 0{,}72\ X_6$$

$$V_2 = 0{,}56\ X_1 + 0{,}42\ X_2 + 0{,}15\ X_3 + 0{,}00\ X_4 - 0{,}02\ X_5 - 0{,}69\ X_6$$

On voit que les variables X_4 et X_5 ne participent pratiquement pas à la détermination des coordonnées des individus dans le plan défini par V_1 et V_2. Ceci s'explique par la petite valeur des quantités

$$\sum_{j=1}^{n} (X_{4j})^2 \quad \text{et} \quad \sum_{j=1}^{n} (X_{5j})^2$$

I_7- b) Pour éviter un tel état de choses nous pouvons faire un changement d'unités de telle sorte que $\sum_{j=1}^{n} (X_{ij})^2 = 1$ pour tout $i = 1, \ldots, 6$.

Ceci revient à choisir

$$\Delta = \begin{vmatrix} \frac{\sqrt{n-1}}{\sqrt{129972{,}80}} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{\sqrt{n-1}}{\sqrt{90474{,}66}} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{\sqrt{n-1}}{\sqrt{26286{,}94}} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{\sqrt{n-1}}{\sqrt{27{,}31}} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{\sqrt{n-1}}{\sqrt{899{,}34}} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{\sqrt{n-1}}{\sqrt{248138{,}58}} \end{vmatrix}$$

La matrice obtenue après ce changement de variables est la suivante :

$R =$

1					
0,98	1				
0,89	0,92	1			
0,77	0,79	0,79	1		
0,67	0,71	0,75	0,65	1	
0,56	0,61	0,66	0,60	0,93	1

La trace D est égale à 6 et les valeurs propres sont les suivantes :

Valeur propre	4,774	0,757	0,290	0,106	0,057	0,016
Participation à D	79,6 %	12,6 %	4,8 %	1,8 %	0,09 %	0,03 %
cumulée	79,6 %	92,2 %	97 %	98,8 %	99,7 %	100 %

On voit qu'ici il faut attendre quatre composantes principales pour absorber un pourcentage de D équivalent au pourcentage absorbé par les deux premières composantes de I_7- a).

Les deux premières composantes sont :

$$W_1 = 0,42\ X_1 + 0,43\ X_2 + 0,43\ X_3 + 0,39\ X_4 + 0,40\ X_5 + 0,37\ X_6$$

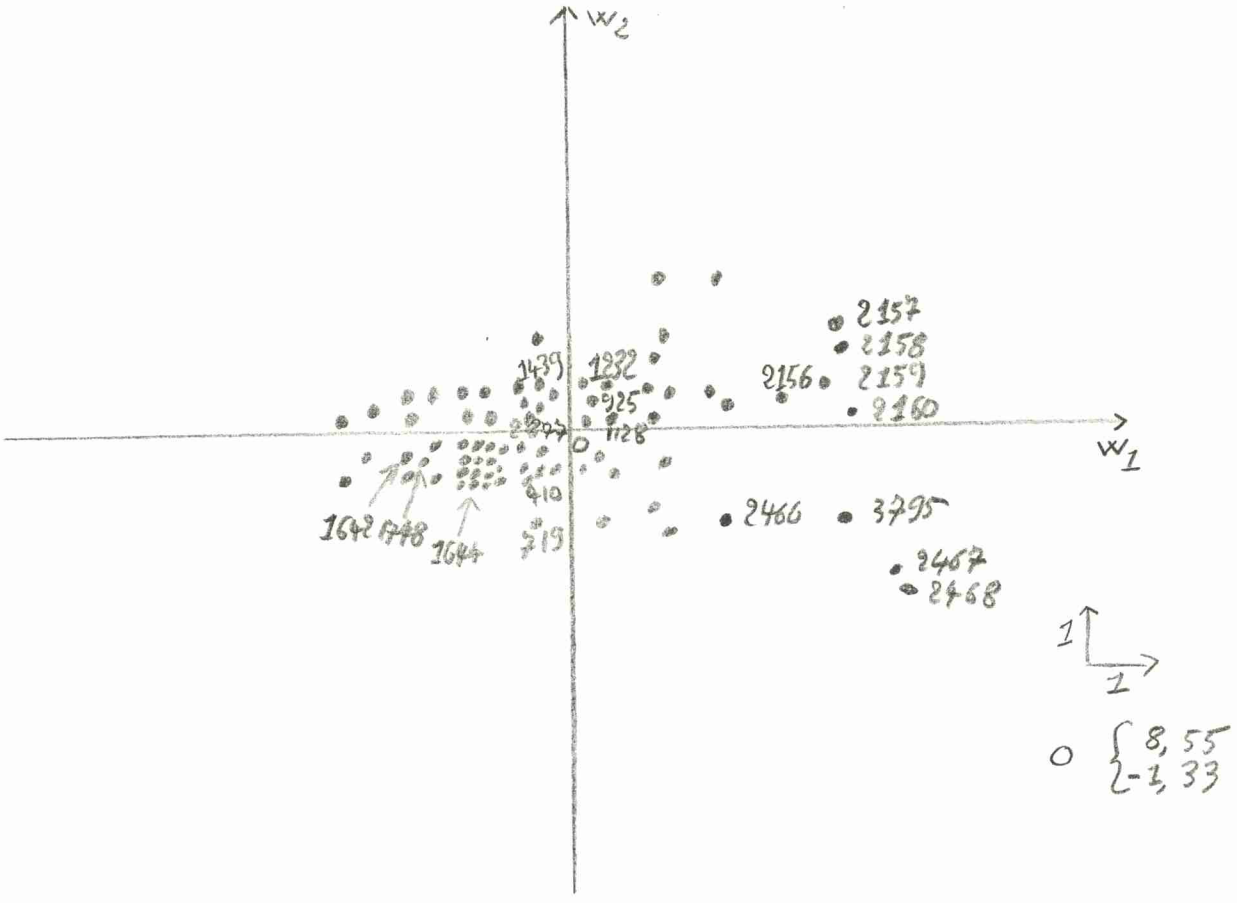
$$W_2 = -0,37\ X_1 - 0,32\ X_2 - 0,18\ X_3 - 0,19\ X_4 + 0,52\ X_5 + 0,65\ X_6$$

On notera qu'ici la participation de X_4 et X_5 est du même ordre que celle des autres variables et que les signes sont changés entre V_2 et W_2. A la symétrie près par rapport à W_1 (due au changement de signe) la représentation sur W_1, W_2 est semblable à celle sur V_1, V_2 mais en plus condensée (figure page 10)

Cette similitude n'est pas générale : elle s'explique ici par le fait que les variations de X_4 et X_5 se font dans le même sens que celles des autres variables mais il n'est pas utopique d'imaginer un exemple dans lequel des variables qui auraient une participation très faible à la trace de Λ seraient croissantes là où les autres seraient décroissantes.

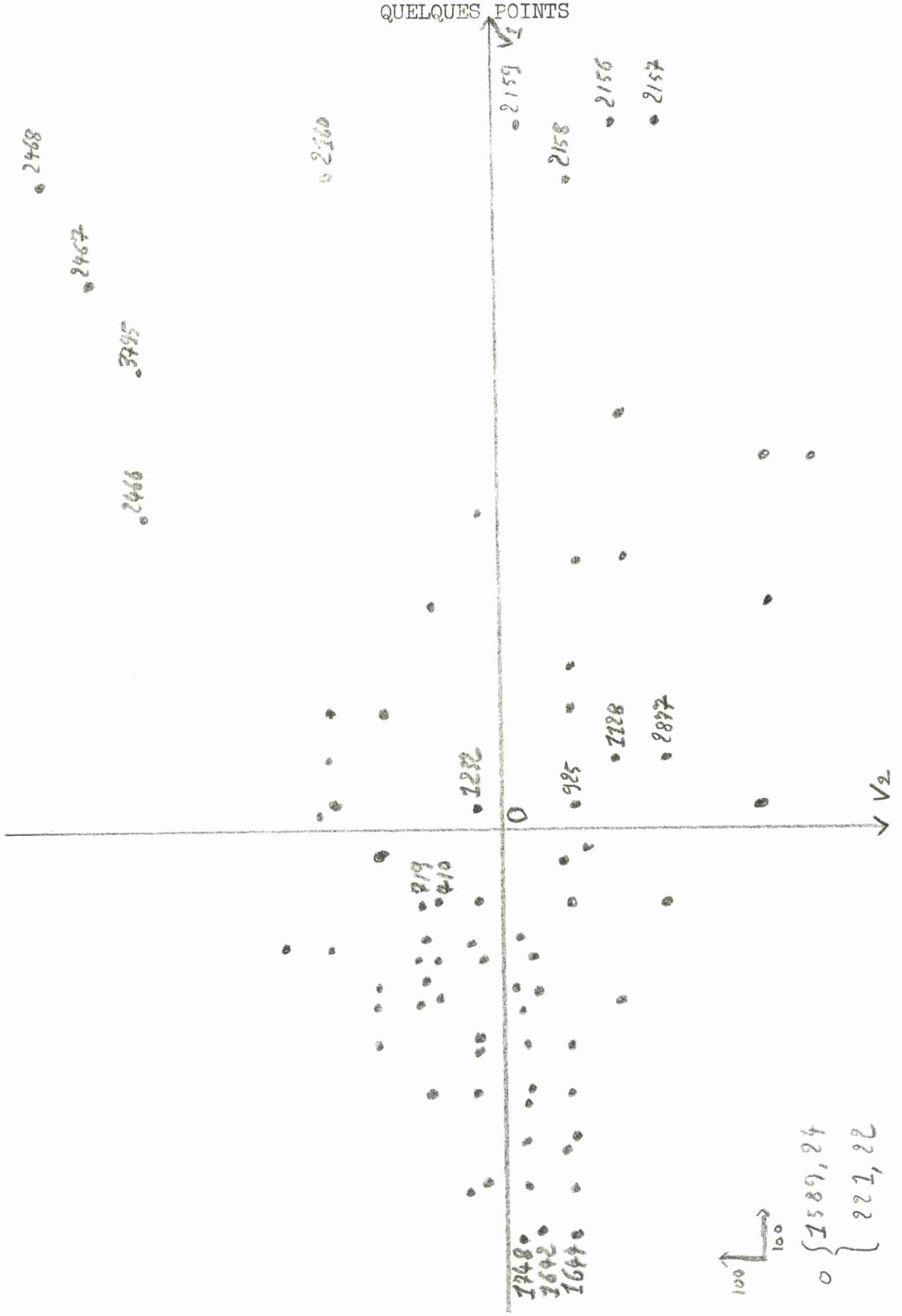
REPRESENTATION OBTENUE A PARTIR DE LA MATRICE R

ET IDENTIFICATION DE QUELQUES POINTS



Les numéros sont les identificateurs des relevés

REPRESENTATION OBTENUE A PARTIR DE LA MATRICE Λ^* ET IDENTIFICATION DE QUELQUES POINTS



CHAPITRE II

II_1- En général, le statisticien conçoit les mesures qu'il possède comme un échantillon tiré d'une population plus grande et entend, à partir des résultats obtenus, inférer des résultats plus généraux valables pour la population.

Dans une telle démarche, l'hypothèse de normalité des variables s'introduit de façon impérative car elle commande l'utilisation des tests de signification qui devront être faits.

Il ne faut pas voir dans l'introduction de l'hypothèse de normalité des variables qu'une condition restrictive susceptible d'interdire l'application de la théorie dans de nombreux cas. En effet, lorsqu'elle est satisfaite, cette hypothèse permet de poser le problème sur des bases nouvelles qui, nous le verrons, permettent des développements plus riches.

Dans les paragraphes II_2 et II_3 nous présentons rapidement la théorie des composantes principales telle qu'elle est généralement développée pour une population multi-normale. Les paragraphes II_4 et II_5 se placent dans l'optique de l'échantillonage et justifient l'utilisation la plus fréquente de la théorie.

II_2- Supposons donc les p variables, notées sur les individus d'une population, distribuées normalement. Dans la représentation géométrique qui est proposée au paragraphe I_1, les lieux des points où la densité de probabilité est uniforme sont des ellipsoïdes concentriques, semblables, dont les axes principaux sont portés par les mêmes droites, qui ont toutes pour centre le point moyen et pour équation

$$X' \Lambda^{-1} X = C$$

où C est une constante, Λ^{-1} l'inverse de la matrice des co-variances de la population et $X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}$ le vecteur observation. ([8] p. 252)

Par analogie avec le problème de la recherche des directions privilégiées du chapître I, notre but est ici de déterminer la direction des axes principaux de ces ellipsoïdes et leur longueur pour une valeur de la constante C.

Soit M un point de l'ellipsoïde de coordonnées $(X_1,\ldots,X_p)$.

Le carré de la distance de M au centre de l'ellipsoïde est $D^2 = \sum_{i=1}^{p} X_i^2$. Si M se déplace continuement sur l'ellipsoïde de manière à la parcourir en totalité, il y aura, pour D^2, p groupes de 2 extremas placés symétriquement par rapport au centre. Ces p groupes de 2 points définissent les p axes principaux de l'ellipsoïde.

Pour trouver ces p axes principaux nous devons chercher les extrémas de $\sum_{i=1}^{p} X_i^2$ avec pour condition $X' \Lambda^{-1} X = C$.

La méthode du multiplicateur de Lagrange nous conduit à égaler à zéro les dérivées par rapport aux X_i de :

$$2T = \sum_{i=1}^{p} X_i^2 + 2\lambda (C - X'\Lambda^{-1}X)$$

d'où $II_{2.1}$ $\qquad X - \lambda\Lambda^{-1} X = 0$

et en multipliant à gauche par Λ

$II_{2.2}$ $\qquad \Lambda X - \lambda X = 0$

Remarque : Nous utiliserons plus loin le résultat sous la forme transposée :

$II_{2.1'}$ $\qquad X' - \lambda X' \Lambda^{-1} = 0$

$II_{2.2'}$ $\qquad X'\Lambda - \lambda X' = 0$

On voit alors que les solutions du problème sont les vecteurs propres de la matrice Λ des variances et co-variances que l'on sait être

définie positive ([3] p.19). Nous individualiserons ces vecteurs propres et les valeurs propres correspondantes que nous supposerons différentes deux à deux, en les indiçant de 1 à p.

On tire facilement de $II_{2.1}$ que $X'_i X_i = \lambda_i C$ et $X'_i X_k = 0$ ([8] p. 254).

Convenons de noter V_i le vecteur propre normé associé à λ_i pour $C = 1$.

Nous appellerons V la matrice dont la $i^{ième}$ colonne est le vecteur V_i. La matrice V est orthogonale $V' = V^{-1}$.

Remarque : Si deux valeurs propres sont égales, la section de l'ellipsoïde par le plan défini par les vecteurs propres associés est un cercle, c'est-à-dire que ces vecteurs ne sont pas définis de manière unique mais qu'il est possible de les choisir d'une infinité de façon dans ce plan. On peut, en particulier les choisir orthogonaux si bien que rien n'est changé dans la définition de la matrice V. Ce résultat se généralise au cas où $q > 2$ valeurs propres sont égales.

II_3- Nous définirons la $i^{ième}$ composante principale comme la projection du vecteur observation sur le vecteur V_i : $Y_i = V'_i X$ $II_{3.1}$

Propriétés des composantes principales

a) $E(Y'_i Y_k) = E\left[(V'_i X)(X' V_k)\right] = V'_i E(XX')V_k$ $II_{3.2}$

$E(XX')$ est par définition la matrice Λ des variances et co-variances de la population. On a donc :

$$E(Y'_i Y_k) = V'_i \Lambda V_k = (V'_i V_k) \lambda_k \qquad II_{3.3}$$

ou bien $i = k$ et $E(Y_i)^2 = \lambda_i$

ou bien $i \neq k$ et $E(Y'_i Y_k) = 0$

La co-variance de deux composantes principales est nulle, ce qui établit

que les composantes principales sont indépendantes puisqu'elles sont distribuées normalement ([3] p16) comme combinaison linéaire de variables distribuées normalement par hypothèse ([1] p.19).

La variance de la $i^{ième}$ composante principale est égale à la $i^{ième}$ valeur propre de la matrice des variances et co-variances.

b) Posons $Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_p \end{bmatrix}$. D'après la définition de V et $II_{3.1}$:

$$Y = V'X$$

et par conséquent $X = VY$ $\quad II_{3.4}$

Proposons-nous de calculer la co-variance de X_k et Y_i.

D'après $II_{3.4}$

$$E(X_k Y_i) = E\left[(v_{1k} Y_1 + v_{2k} Y_2 + \ldots + v_{pk} Y_p)(Y_i)\right]$$

$$E(X_k Y_i) = \sum_{j=1}^{p} v_{jk} \; E(Y_j Y_i)$$

La propriété précédente établit alors que :

$$E(X_k Y_i) = v_{ik} \lambda_i \qquad II_{3.5}$$

II_4- Le résultat important, qui justifie l'utilisation de la méthode telle qu'elle est pratiquée couramment à partir d'un échantillon, est que l'estimation, au sens du maximum de vraisemblance, des composantes principales et de leur variance, est donnée par les vecteurs propres et les valeurs propres de la matrice estimée des variances et covariances. Je ne démontrerai pas ce résultat que l'on trouvera dans Anderson ([1] p.279).

Nous pouvons alors nous demander à juste titre si les différences que nous voyons apparaître entre les valeurs propres de la matrice estimée sont indicatrices de différences effectives entre les valeurs propres de la matrice véritable, ou bien si elles ne sont dues qu'aux erreurs d'échantillonage.

En effet, supposons que nous ayons obtenu l'estimation des k plus grandes valeurs propres, l'hypothèse de l'égalité des p-k restantes si elle est vraie, implique que toutes les variables aléatoires de la forme $Y = \sum_{i=1}^{p} v_i X_i$, telles que $\sum_{i=1}^{p} v_i^2 = 1$, et non corrélées avec $Y_1, Y_2, \ldots, Y_k$ ont la même variance.

Il est donc inutile de chercher à rendre maximum cette variance et l'analyse s'arrête.

Dans le cas où les k premières composantes principales ont absorbées une grand part de la variation, cela revient à se demander si Y ne peut pas être partagé en un terme $Y^* = \begin{bmatrix} Y_1 \\ \vdots \\ Y_k \end{bmatrix}$ qui représente les effets réels et un terme d'erreur $Y^{**} = \begin{bmatrix} Y_{k+1} \\ \vdots \\ Y_p \end{bmatrix}$, orthogonal à Y^*, mais qui, cette orthogonalité respectée n'a pas de direction privilégiée.

Le procédé utilisé pour tester l'égalité des p-k plus petites valeurs propres lorsqu'on a estimé les k premières est du à Bartlett ([10])

Soit Λ^* l'estimation de la matrice des variances et co-variances et $|\Lambda^*|$ son déterminant. Si n est le nombre d'individus que compte l'échantillon, pour tester si les différences entre les p-k plus petites valeurs propres sont significatives, on utilise la quantité :

$$\top = n' \{-\log_e |\Lambda^*| + \log_e (\lambda_1 \ldots \lambda_k) + (p-k) \log_e \lambda \}$$

où

$$\lambda = \{tr\, \Lambda^* - \lambda_1 - \lambda_2 \ldots - \lambda_k\} \times \frac{1}{p-k}$$

et $n' = n - k - \frac{1}{6}\left[2(p-k)+1+\frac{2}{p-k}\right]$

Cette quantité est distribuée asymptotiquement comme un χ^2 à $\frac{1}{2}(p-k+2)(p-k-1)$ degrés de liberté :

Exemple : Considérons les données de I_7 a) comme constituant un échantillon. L'hypothèse de normalité étant tout à fait légitime pour les 6 variables retenues, nous nous trouvons dans le cadre prévu pour pouvoir appliquer les résultats que nous venons d'expliciter.

La matrice Λ^* de I_7 a), telle qu'elle a été construite, est la matrice des variances et co-variances estimée à partir de l'échantillon.

Les deux premières valeurs propres ayant été estimées, nous voulons tester l'égalité des quatre plus petites valeurs propres dans la population.

Le calcul donne : $T = 15,83$

Le nombre d'individus est 76. Nous comparons donc T au χ^2 donné par la table à 9 degrés de liberté, qui est égal à 16,92 pour un seuil de signification de 5 %. Il n'y a donc pas lieu de rejeter l'hypothèse de l'égalité des plus petites valeurs propres.

Si au contraire, après avoir extrait la plus grande valeur propre, nous avions testé l'égalité des cinq plus petites, nous aurions trouvé $T = 49,14$ pour 14 degrés de liberté et nous aurions dû rejeter l'hypothèse.

II_5- Il est très rare de trouver dans les publications une analyse des composantes principales faite à partir de la matrice des variances et co-variances. Les statisticiens préfèrent travailler sur la matrice des corrélations.

Dans sa présentation de la théorie, Hotelling arrive à la matrice des corrélations car il suppose que les variables sont réduites et dans ce cas, bien sûr, la matrice des corrélations et la matrice des co-variances s'identifient.

Il faut bien voir que, lorsque dans le cas général le statisticien décide de travailler sur la matrice des corrélations, il décide en fait de réduire les variables, c'est-à-dire de donner à chacune la même part dans la dispersion totale.

Cette hypothèse n'est pas essentielle à la théorie. Il peut être tout à fait justifier de donner à certaines variables une participation plus grande qu'à d'autres, soit en travaillant sur la matrice des covariances, soit en affectant certaines variables d'un poids.

Le résultat de l'analyse est lié au choix des unités dans lesquelles sont exprimées les différentes variables, c'est-à-dire à la définition que l'on se donne de la distance entre deux individus (à la métrique dont on munit l'espace).

Il peut être normal d'attribuer à certaines variables une part primordiale dans la définition de cette distance si bien que l'utilisation de la matrice des corrélations doit être moins automatique qu'elle ne l'est souvent.

Remarque 1 : Lorsqu'on travaille avec les variables réduites, la conjonction des formules $II_{3.3}$ et $II_{3.5}$ donne

$$r(X_k\ Y_i) = v_{ik}\sqrt{\lambda_i} \qquad II_{5.1}$$

où $r(X_k Y_i)$ est le coefficient de corrélation entre X_k et Y_i

Remarque 2 : Il existe un procédé, dû à Bartlett, pour tester l'égalité des p-k plus petites valeurs propres de la matrice des corrélations lorsque l'on a estimé les k premières.

Le critère est $T = n \{-\log_e |R| + \log_e (\lambda_1 \ldots \lambda_k) + (p-k) \log_e (\lambda)\}$

où $\lambda = (p-\lambda_1 \ldots - \lambda_k) \times \frac{1}{p-k}$

Malheureusement, même à la limite ce critère ne suit pas exactement une distribution de χ^2. Toutefois, si les valeurs propres éliminées représentent une très grande proportion de la variance totale, l'espérance mathématique de T qui devrait être prise comme le nombre de degrés de liberté du χ^2 est voisine de $\frac{1}{2}(p-k+2)(p-k-1)$. [9]

Exemple : Telle qu'elle a été construite en I_7 b), la matrice R est la matrice des corrélations estimées à partir de l'échantillon. La remarque 1 nous permet de construire le tableau des corrélations entre les composantes principales, et les variables.

	x_1	x_2	x_3	x_4	x_5	x_6
W_1	0,92	0,94	0,94	0,86	0,87	0,81
W_2	- 0,32	- 0,28	- 0,16	- 0,16	0,45	0,57

On remarquera la valeur élevée des coefficients de corrélations entre les variables et la première composante principale. Des résultats semblables seront interprétés au chapitre suivant.

Après l'obtention des trois premières valeurs propres, on peut tester l'hypothèse de l'égalité des trois suivantes par le test de Bartlett. On trouve $T = 59,83$ pour 5 degrés de liberté ce qui nous conduit à rejeter l'hypothèse. Rappelons les doutes émis sur la validité de ce test.

CHAPITRE TROISIEME

III-1 Ce chapitre voudrait établir qu'une organisation judicieuse de la matrice de corrélations, permet de mettre en évidence la signification des composantes principales et, dans certains cas, est susceptible d'en donner, sans calcul, une excellente approximation.

Les données traitées ci-dessous sont extraites d'une étude sylvicole effectuée dans des pineraies de Sologne par M. Godron, Chef de la Section de recherche fondamentale au Centre d'Etudes Phytosociologiques et Ecologiques de Montpellier.

Sur cent et une parcelles sylvicoles, on a noté douze variables dont voici la liste.

Variable 1 : Recouvrement de l'étage dominant.

Variable 2 : Epaisseur du A_o

Variable 3 : Profondeur engorgée en hiver

Variable 4 : Profondeur de l'Argile

Variable 5 : Activité biologique

Variable 6 : Humidité déduite de la flore

Variable 7 : Age

Variable 8 : Indice de productivité (hauteur ramenée à 50 ans)

Variable 9 : Nombre d'arbres à l'hectare.

Variable 10 : Surface terrière moyenne.

Variable 11 : Volume de bois à l'hectare.

Variable 12 : Hauteur moyenne.

Les variables 1,5,6 sont codées selon un protocole établi par le C.E.P.E.

Les variables 2,3,4 sont exprimées en cm.

Les variables 8,12 sont exprimées en m.

La variable 10 en m^2.

La variable 11 en m^3.

La variable 7 en année.

La diversité des unités employées conduit tout naturellement, d'après II_5, à utiliser la matrice des corrélations, c'est-à-dire à réduire les variables.

L'hypothèse de normalité pour ces douze variables est approximativement confirmée par les histogrammes de fréquence construits à partir des cent un relevés.

On trouvera la matrice des corrélations en page 22 et ses valeurs propres et ses vecteurs propres en page 23.

Portons notre attention sur les coordonnées des composantes principales.

La première composante a sur les variables 7,9,10, 11, 12 des coordonnées qui, en valeur absolue, sont comprises entre 0,3665 et 0,4845 alors que, sur les autres variables, les coordonnées sont beaucoup plus faibles. D'autre part, les variables 7, 9, 10, 11, 12 ont entre elles de fortes corrélations (mis à part 9 et 10) qui dépassent de beaucoup les corrélations entre l'une quelconque des variables 7, 9, 10, 11, 12 et les variables restantes.

Nous pouvons faire la même constatation avec la deuxième composante et les variables 1, 3, 4, 5, 6, 8 si bien qu'il apparaît que les deux composantes les plus importantes sont liées à des groupes de variables fortement corrélées entre elles. Peut-on préciser ce résultat ?

III_2- Considérons la matrice X des observations contrées, introduite en I_1, non plus comme donnant les p coordonnées de n points dans

MATRICE DES CORRELATIONS

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
x_1	1											
x_2	-0,0528	1										
x_3	0,4039	0,0471	1									
x_4	0,3018	-0,1841	0,6445	1								
x_5	0,2566	-0,4080	0,5237	0,4316	1							
x_6	0,2848	0,1588	0,6578	0,4163	0,5395	1						
x_7	-0,3084	0,2535	0,0152	-0,0675	-0,0912	-0,0796	1					
x_8	0,2245	-0,1824	0,2559	0,1579	0,5533	0,2180	-0,2450	1				
x_9	0,3544	-0,1975	0,0837	0,2170	0,0924	0,1207	-0,7508	0,1736	1			
x_{10}	0,0259	0,2309	0,2511	0,1039	0,1996	0,1317	0,5591	0,3091	-0,1857	1		
x_{11}	-0,1385	0,2564	0,1853	0,0226	0,1460	0,0433	+0,7895	0,2395	-0,5368	0,8921	1	
x_{12}	-0,1989	0,2148	0,1160	-0,0510	0,1274	0,0026	0,8872	0,1796	-0,7402	0,6906	0,9158	1

VALEURS PROPRES ET VECTEURS PROPRES

Valeur propre	4,0012	3,2467	1,3330	1,0532	0,7614	0,6257	0,4216	0,2396	0,1752	0,1068	0,0239	0,0117
cumulée	4,0012	7,2479	8,5809	9,6341	10,3955	11,0212	11,4428	11,6824	11,8576	11,9644	11,9883	12,0000
pourcentage	33,3 %	60,4 %	71,5 %	80,3 %	86,6 %	91,8 %	95,4 %	97,4 %	98,8 %	99,7 %	99,9 %	100 %
	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}	Y_{11}	Y_{12}
x_1	-0,1338	0,3036	0,1599	-0,2956	-0,4609	0,7295	-0,0694	**0,1606**	0,0171	**0**,0061	-0,0102	0,0375
x_2	0,1676	-0,1027	0,6272	-0,4299	0,2780	-0,0638	0,2742	0,1732	-0,4381	-0,0641	0,0238	0,0068
x_3	0,0584	0,4554	0,2991	0,1614	-0,0317	-0,0057	0,1667	-0,7994	-0,0351	-0,0616	0,0286	0,0003
x_4	-0,0331	0,3891	0,1573	0,3526	-0,4071	-0,3747	0,4270	0,4510	0,0055	0,0755	-0,0021	-0,0315
x_5	0,0232	0,4440	-0,3145	0,2097	0,2669	0,1015	-0,2587	0,1454	-0,6997	-0,0113	-0,0073	0,0063
x_6	0,0083	0,3992	0,3507	0,0943	0,5031	0,0108	-0,3632	0,2569	0,5054	0,0262	-0,0378	0,0047
x_7	0,4559	-0,0977	0,0828	0,2350	-0,1562	0,0549	-0,1584	0,0838	0,0194	-0,5384	0,3948	0,4636
x_8	0,0356	0,3253	-0,4646	-0,4120	0,2858	0,0151	0,5082	0,0354	0,2259	-0,1607	0,2068	0,2150
x_9	-0,3665	0,1637	-0,0067	-0,3697	-0,2086	-0,4338	-0,3238	-0,0271	-0,0340	-0,5752	-0,1701	-0,0628
x_{10}	0,3813	0,1760	-0,0625	-0,3695	-0,2357	-0,2977	-0,3293	-0,0438	0,0078	0,4263	0,4608	-0,1859
x_{11}	0,4746	0,0901	-0,0846	-0,1586	-0,1182	-0,1040	-0,0635	-0,0243	0,0302	0,1532	-0,7237	0,3958
x_{12}	0,4845	0,0280	-0,1059	0,0356	0,0006	0,1489	0,1016	+0,0565	0,0893	-0,3657	-0,1826	-0,7357

un espace à p dimensions, mais les n coordonnées de p points dans un espace à n dimensions, supposé rapporté à des axes orthogonaux, d'origine O.

Dans cet espace, les p variables sont représentées par p points $X_1, \ldots, X_p$ ou par p vecteurs ayant O pour origine et $X_1, \ldots, X_p$ pour extrémités, les coordonnées de X_i étant $X_{i1}, X_{i2}, \ldots, X_{in}$.
Supposons donné dans l'espace à n dimensions le produit scalaire habituel et la norme qui en découle.
Il est aisé de voir que le cosinus de l'angle θ_{ij}, des vecteurs $\overrightarrow{OX}_i$ et $\overrightarrow{OX}_j$ est égal au coefficient de corrélation des variables x_i et x_j.

En effet : $\overrightarrow{OX}_i \ \overrightarrow{OX}_j = \|OX_i\| \ \|OX_j\| \cos \theta_{ij}$

d'où : $$\cos \theta_{ij} = \frac{\overrightarrow{OX}_i \ \overrightarrow{OX}_j}{\|OX_i\| \ \|OX_j\|} = \frac{\sum_{k=1}^{n} X_{ik} X_{jk}}{\sqrt{\left(\sum_{k=1}^{n} X_{ik}^2\right)\left(\sum_{k=1}^{n} X_{jk}^2\right)}} = r(x_i, x_j)$$

Donc, dans cet espace, des variables fortement corrélées entre elles, sont représentées par des vecteurs qui rayonnent autour de l'origine comme les génératrices d'un cône dont l'angle au sommet est d'autant plus petit que les corrélations sont fortes.

Nous pouvons représenter les composantes principales dans cet espace
Soit $r(x_k, Y_i)$ le coefficient de corrélation entre la variable x_k et la composante Y_i. D'après $II_{5.1}$

$$r(x_k, Y_i) = v_{ik} \sqrt{\lambda_i}$$

Posons $\rho = \max_k \left(|r(x_k, Y_i)| \right) \leq 1$. On a alors :

$$|v_{ik}| \leq \frac{\rho}{\sqrt{\lambda_i}} \leq \frac{1}{\sqrt{\lambda_i}}$$

ce qui établit que les coefficients v_{ik} sont d'autant plus petits que la valeur propre est grande.

Les v_{ik} ont été calculés de telle sorte que $\sum_{k=1}^{p} v_{ik}^2 = 1$ si bien que, pour i donné, le nombre des coefficients v_{ik} non négligeables varie comme la taille de la valeur propre λ_i, ou, pour revenir à la représentation dans l'espace à n dimensions, le nombre de variables faisant un angle petit avec la composante principale est liée à la taille de la valeur propre associée à cette composante. Mais, si des variables font un angle petit avec une composante principale, elles font un angle petit entre elles ce qui nous conduit à dire que des valeurs propres fortes sont liées à l'existence de groupes de variables fortement corrélées entre elles.

III$_3$- Ceci nous conduit à envisager la recherche des groupes de variables comme un préalable à l'analyse des composantes principales.

S'intéressant à cette question Monsieur R. Tomassone (Communication verbale et [21]) propose un moyen lié à la théorie des graphes. Les variables sont placées aux sommets du graphe. Choisissant un seuil (arbitraire) de signification pour les coefficients de corrélations, on joint par un trait les variables qui présentent des corrélations supérieures à ce seuil. Par abaissement du seuil (éventuellement sanctionné par un changement de couleur pour les arêtes), on introduit progressivement les variables. Le problème de la recherche des groupes est alors ramené à celui de la recherche des parties connexes d'un graphe.

Cherchant un moyen plus systématique et susceptible d'être programmé afin que son utilisation puisse être rapide, j'ai utilisé le procédé suivant :

a) Les deux variables qui présentent le plus fort coefficient de corrélation (en valeur absolue) sont extraites de la matrice et donnent naissance au premier groupe.

b) La variable dont la somme des coefficients de corrélation (en valeur absolue) avec les variables déjà choisies est maximum est isolée.

c)-Ou bien cette variable peut être considérée comme faisant partie du groupe des variables déjà choisies : elle y est intégrée et on recommence en b).

-ou bien cette variable ne peut pas être considérée comme faisant partie du groupe qui est alors clos. La variable rejetée est réintégrée à la matrice et l'on recommence en a) avec les variables non encore choisies.

La difficulté repose dans le test qui doit décider de l'appartenance au groupe. H.H. Harman [16] dans sa présentation de la méthode d'analyse factorielle appelée "Group-Factor solution" propose un coefficient B très difficilement applicable car l'appréciation de la signification de la différence de deux valeurs successives de ce coefficient est tout à fait arbitraire.

J'ai utilisé un procédé qui, bien que ne présentant pas toutes les assurances statistiques souhaitables, a l'avantage de donner naissance à un programme rapide et de mener à des résultats satisfaisants.

Supposons que le groupe soit déjà constitué de k variables ($k \geq 1$) et introduisons la $(k+1)^{ième}$ variable (choisie selon b).

Aux $k(k+1)/2$ coefficients de corrélations, faisons subir la transformation $Z = \frac{1}{2}\left[\text{Log } (1+r) - \text{Log } (1-r) \right]$ et calculons la valeur moyenne, $\overline{Z}$ ([7] p.175). Prenant pour estimation de la variance de $\overline{Z}$,

la quantité $\sqrt{\frac{1}{n-3}}$, la valeur $\overline{Z} + t\sqrt{\frac{1}{n-3}}$ est considérée comme une valeur maximum d'estimation qui varie avec le seuil de probabilité choisi pour t. Ensuite, si un au moins des coefficients de corrélation de la $(k+1)^{ième}$ variable avec les variables restantes donne $Z > \overline{Z} + t\sqrt{\frac{1}{n-3}}$, la $(k+1)^{ième}$ variable est considérée comme ne faisant pas partie du groupe.

Après un tel classement, la matrice des corrélations s'organise de la manière suivante :

$$\tilde{R} = \begin{pmatrix} R_{11} & R_{12} \cdots & R_{1k} \\ R_{21} & R_{22} & \\ \vdots & \ddots & \\ R_{k1} & & R_{kk} \end{pmatrix}$$

où les R_{ii} sont des matrices symétriques dont les éléments sont des coefficients de corrélation élevées (en valeur absolue) et les R_{ij} des matrices dont les éléments sont des coefficients de corrélations faibles.

Pour avancer notre problème, il reste à montrer que l'on peut localiser les valeurs propres de $\tilde{R}$ à partir des valeurs des R_{ii} et que l'on peut déduire, dans une bonne approximation, les vecteurs propres de $\tilde{R}$ des vecteurs propres des R_{ii}.

III_4- Dans ce but rappelons quelques résultats d'analyse numérique ([24] p.275).

III_4- a) Soit A une matrice symétrique, X_i les vecteurs normés, λ_i les valeurs propres supposées distinctes. Par définition

$$AX_i = \lambda_i X_i \qquad III_{3.1}$$

Supposons que les éléments de A subissent une légère pertubation représentée par la matrice δA. Le $i^{ième}$ vecteur propre (non

normé) de $A + \delta A$ peut être pris égal à $X_i + \sum_{j \neq i} \alpha_{ij} X_j$ et la $i^{ième}$ valeur propre à $\lambda_i + \delta\lambda_i$, si bien que :

$$(A + \delta A)(X_i + \sum_{j \neq i} \alpha_{ij} X_j) = (\lambda_i + \delta\lambda_i)(X_i + \sum_{j \neq i} \alpha_{ij} X_j) \quad III_{3.2}$$

Convenons de négliger les termes du deuxième ordre et tenons compte de $AX_j = \lambda_j X_j$, alors :

$$\delta A X_i + \sum_{j \neq i} \alpha_{ij} \lambda_j X_j = \delta\lambda_i X_i + \sum_{j \neq i} \alpha_{ij} \lambda_i X_j \quad III_{3.3}^{W}$$

En multipliant à gauche par X'_i et en tenant compte de $X'_i X_u = \delta_{ij}$ (Symbole de Krönecker), on obtient

$$X'_i \, \delta A \, X_i = \delta\lambda_i$$

d'où :

$$|\delta\lambda_i| \leq \|\delta A\| \quad III_{3.4}$$

Ce qui établit que de faibles variations dans les éléments d'une matrice symétrique entraînent de faibles **variations** des valeurs propres.

III_4- b) En multipliant à gauche $III_{3.3}$ par X'_k $(k \neq i)$, on obtient

$$X'_k \, \delta A X_i + \alpha_{ik} \lambda_k = \alpha_{ik} \lambda_i$$

ou :

$$\alpha_{ik} = \frac{X'_k \delta A X_i}{\lambda_k - \lambda_i}$$

Soit

$$|\alpha_{ik}| \leq \frac{\|\delta A\|}{|\lambda_k - \lambda_i|} \quad III_{3.5}$$

Si les différences entre les valeurs propres **sont importantes**, les α_{ik} seront donc petits et les vecteurs propres peu **perturbés**. Au contraire, si les valeurs propres sont voisines, de grandes perturbations sur les vecteurs propres peuvent être attendues.

III_4- c) Supposons une matrice A symétrique $p \times p$ dont tous les éléments diagonaux soient égaux à 1, et tous les éléments non diago-

naux à r ([5] p. 13)

En ajoutant toutes les lignes à la première, en mettant en facteur le terme $(1 + (p-1)r)$ et en soustrayant la dernière colonne de toutes les autres, il est facile de voir que le déterminant est égal à

$$(1-r)^{p-1}(1+(p-1)r)$$

si bien qu'une telle matrice a une grande valeur propre égale à $1+(p-1)r$ et $p-1$ petites valeurs propres égales à $(1-r)$.
Il est immédiat que le vecteur propre associé à la plus grande valeur propre est défini par :

$$v_1 = v_2 = \ldots = v_p$$

III_4- d) Supposons enfin que A, matrice symétrique $p \times p$, soit de la forme :

$$A = \begin{pmatrix} A_{ii} & 0 \\ 0 & A_{22} \end{pmatrix}$$

où A_{ii} et une matrice symétrique $p_i \times p_i$ $(p_1 + p_2 = p)$
Nous cherchons les vecteurs W tels que :

$$AW = \lambda W \qquad III_{3.6}$$

Décomposons W en un vecteur W_1 à p_1 dimensions et W_2 à p_2 dimensions ; $III_{3.6}$ devient

$$\begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix} \begin{pmatrix} W_1 \\ W_2 \end{pmatrix} = \lambda \begin{pmatrix} W_1 \\ W_2 \end{pmatrix}$$

ou

$$\begin{cases} A_{11} W_1 = \lambda W_1 \\ A_{22} W_2 = \lambda W_2 \end{cases} \qquad III_{3.7}$$

ou bien A_{11} et A_{22} ont des valeurs propres en commun qui seront des valeurs propres d'ordre 2 de A ; ou bien A_{11} et A_{22} n'ont pas de valeurs propres communes et les valeurs propres de A seront les valeurs propres μ_i de A_{11} associées aux vecteurs propres

$\begin{pmatrix} V_i \\ 0 \end{pmatrix}$ où V_i est le vecteur propre de A_{11} associé à μ_i, et les valeurs propres ν_k de A_{22} associé aux vecteurs propres $\begin{pmatrix} 0 \\ U_k \end{pmatrix}$

où U_k est le vecteur propre de A_{22} associé à ν_k.

III_5- La conjonction des résultats du paragraphe précédent va alors permettre de prévoir à l'avance la taille des valeurs propres importantes qui apparaîtrons dans notre étude et la forme des vecteurs propres attendues. D'autre part, en isolant ainsi dès l'origine les variables qui vont participer à la définition d'une composante principale, la signification de cette dernière est plus facile à saisir.

Reprenons l'exemple proposé en III_1. Le classement effectué par le procédé défini en III_3 fait apparaître deux groupes de variables

1er groupe : 11, 12, 7, 10, 9

2ième groupe : 3, 6, 5, 4, 1, 8, 2

La variable essayée à la fin du groupe 1 (et rejetée) est la variable 2. On peut remarquer dans l'optique de la matière traitée que le premier groupe est constitué des variables dites de production, alors que le deuxième groupe englobe les variables de milieu.

Ci-dessous la matrice correspondant au premier groupe

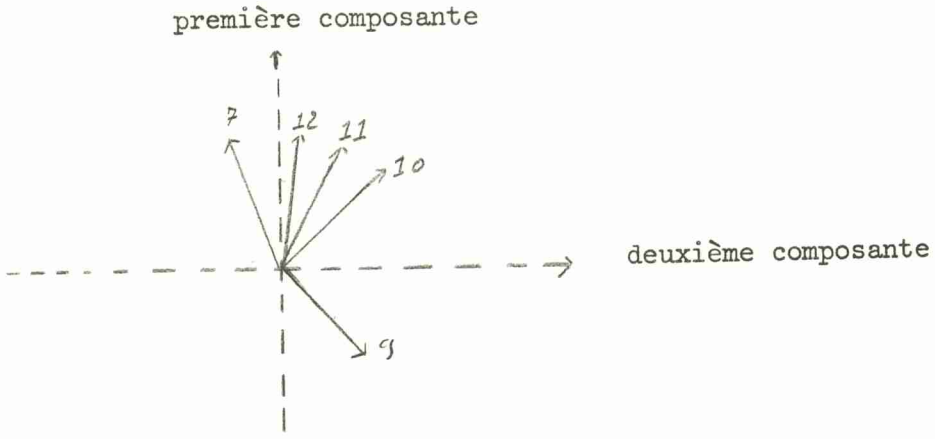
MATRICE DES CORRELATIONS

DU 1er GROUPE

	x_{11}	x_{12}	x_7	x_{10}	x_9
x_{11}	1				
x_{12}	0,9158	1			
x_7	0,7895	0,8872	1		
x_{10}	0,8921	0,6906	0,5591	1	
x_9	-0,5368	-0,7402	-0,7508	-0,1857	1

La moyenne des coefficients de corrélation de cette matrice est 0,70. En utilisant les résultats de continuité du paragraphe III_3-a) et celui du paragraphe III_3-c) on peut prévoir une valeur propre de l'ordre de $(1+4 \times 0,70) = 3,8$.

La faiblesse du coefficient de corrélation entre les variables 9 et 10 nous indique que dans l'espace à n dimensions les vecteurs représentant ces variables ne sont pas loin d'être orthogonaux si bien qu'en tenant compte de l'ensemble de la matrice des corrélations on peut postuler la représentation suivante :



Il est plus difficile de prévoir la taille de la deuxième composante. La représentation graphique suggère que la dispersion due aux variables 9 et 10 se partage également entre les deux composantes, si bien que puisque la participation au total de chaque variable est 1, on peut attendre une deuxième composante de l'ordre de l'unité : ($\frac{1}{2}$ pour x_9 + $\frac{1}{2}$ pour x_{10}).

Le premier vecteur propre devant être normé, III_3- c) nous invite à attendre des v_{1k}^2 de l'ordre de 0,20, soit des v_{1k} de l'ordre de 0,43. La représentation graphique invite à penser que les variables 12, 11, 7 donneront des coefficients supérieurs à ceux de 10 et 9, ce dernier étant négatif.

Pour la deuxième composante, la représentation graphique indique que les plus forts coefficients seront ceux de 9 et 10 ; ceux de 7 et 11 seront de signes opposés et comparables en valeur absolue ; celui de 12 devrait être pratiquement nul.

La somme atteinte des valeurs propres hypothètiques étant voisine de 5, il n'y a pas lieu de s'interroger sur les autres.

Ci-dessous le tableau des valeurs propres et vecteurs propres calculés qui ne contredit pas nos hypothèses.

VALEURS PROPRES ET VECTEURS PROPRES DU 1er GROUPE

valeur propre	3,828	0,924	0,158	0,070	0,020
cumulée	3,828	4,752	4,910	4,980	5
pourcentage	76,5 %	95 %	98,2 %	99,6 %	100 %
	Y_1	Y_2	Y_3	Y_4	Y_5
x_{11}	0,487	0,277	0,209	-0,128	0,791
x_{12}	0,499	-0,067	0,081	-0,747	-0,426
x_7	0,469	-0,224	-0,824	0,220	0,043
x_{10}	0,393	0,647	0,150	0,467	-0,432
x_9	-0,373	0,671	-0,494	-0,398	0,062

Donc pour ce premier groupe, à côté d'une composante de production on voit apparaître une composante de correction qui, constatons-le, le nombre d'arbres à l'hectare étant lié aux coupes qui sont faites, et la surface terrière moyenne aux nombres d'arbres à l'hectare, résume l'intervention de l'homme dans la production.

Remarque : Par le test de Bartlett, après extraction des 2 premières valeurs propres, on trouve T = 94,70 pour 5 degrés de liberté, ce

qui nous amènerait à rejeter l'hypothèse de l'égalité des trois variables restantes.

Nous conviendrons d'appeler composante de groupe, une composante du type de Y_1 à laquelle participent de manière comparable toutes les variables d'un groupe, et composantes secondaires, celles qui comme Y_2, opposent des variables à l'intérieur d'un même groupe.

Prenons maintenant la matrice du deuxième groupe :

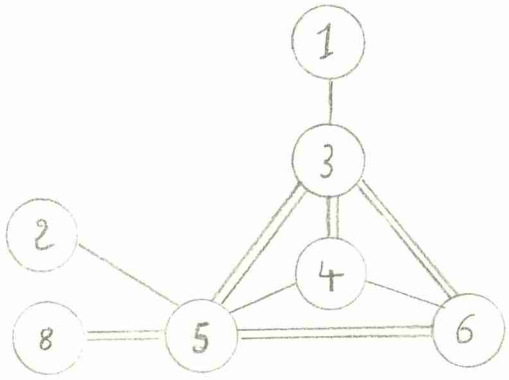
	x_3	x_6	x_5	x_4	x_1	x_8	x_2
x_3	1						
x_6	0,6578	1					
x_5	0,5237	0,5395	1				
x_4	0,6445	0,4163	0,4316	1			
x_1	0,4039	0,2848	0,2566	0,3018	1		
x_8	0,2559	0,2180	0,5533	0,1579	0,2245	1	
x_2	0,0471	0,1588	-0,4080	-0,1841	-0,0528	-0,1824	1

La moyenne des coefficients de corrélations est 0,32, ce qui peut laisser espérer que la composante de groupe soit attachée à une valeur propre de l'ordre de $(1+6 \times 0,32) = 2,92$.

Les coefficients du vecteur propre correspondant devraient être de l'ordre $\sqrt{1/7}$, soit 0,37. Notons cependant que la dispersion des coefficients de corrélation étant grande, nous devons nous attendre à ce que les valeurs réelles ne soient pas très voisines de ces valeurs supposées.

Il est difficile ici de faire la représentation dans l'espace à n dimensions, mais nous pouvons adopter la méthode proposée en III_3. Choisissant arbitrairement comme seuils pour les coefficients de

corrélation 0,50(=) et 0,40(-) , on a :



Cette représentation nous suggère qu'à côté de la composante de groupe qui absorbe la plus grande partie de la variation due aux variables 3, 4, 5, 6, nous allons trouver des composantes secondaires qui absorberont la variation résiduelle non négligeable attachée aux variables 1, 2, 8.

Géométriquement, à côté de la composante de groupe qui définira presque exhaustivement la direction des vecteurs représentant les variables 3,4,5,6, nous trouverons des composantes secondaires nécessaires pour définir l'orientation des vecteurs représentant les variables 1,2,8.

Le calcul des valeurs propres et des vecteurs propres permet d'établir le tableau suivant :

valeurs propres	3,057	1,335	0,869	0,798	0,496	0,252	0,193
cumulée	3,057	4,392	5,261	6,059	6,555	6,807	7
pourcentage	43,6 %	62,7 %	75,1 %	86,5 %	93,6 %	97,2 %	100 %
	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7
x_3	0,480	0,286	-0,115	-0,085	0,133	0,789	-0,167
x_6	0,425	0,346	0,183	-0,276	- 0,499	-0,394	-0,430
x_5	0,465	-0,319	0,150	-0,225	-0,294	0,033	0,722
x_4	0,413	0,084	-0,536	-0,126	0,557	-0,447	0,091
x_1	0,308	0,104	-0,075	0,915	-0,194	-0,089	0,074
x_8	0,305	-0,374	0,669	0,110	0,479	-0,078	-0,268
x_2	-0,120	0,734	0,436	0,006	0,263	-0,083	0,424

En nous limitant par souci de clarté aux composantes principales qui absorbent plus de 10 % de la variation totale, nous obtenons donc :

Y_1 qui est clairement la composante de groupe ;

Y_2 qui oppose les variables 8 et 5 aux autres et surtout à 2.

Y_3 qui oppose les variables 4 (et 3 faiblement) aux autres et surtout à 2 et 8.

Y_4 qui est définie par la variable 1 toute seule avec comme termes correctifs les variables 5 et 6.

Ramenons nous au tableau des valeurs propres et vecteurs propres de la matrice 12 × 12 , donné au début de ce chapître.
Nous reconnaissons facilement dans le premier et le second vecteur de ce tableau, les composantes du premier et du second groupe.
La troisième composante qui oppose les variables 5 et 8 aux variables 2, 3, 6 s'identifie sans difficulté avec la composante Y_2 de la matrice 7 × 7.

Dans la quatrième composante nous trouvons regroupées les oppositions de 9, 10, 11 à 7 (composante Y_2 de la matrice 5 × 5) et de 2, 8 à 4, 5 (composante Y_3 de la matrice 7 × 7)
Le rôle particulier que l'on était en droit d'attendre de x_1 semble s'être reporté sur deux composantes ; dans Y_5, la variable 1 s'oppose aux variables 5 et 6 aux côtés des variables 4, 9, 10, 11 ; en face desquelles elle se trouvera dans Y_6.
L'intérêt de cette étude analytique des composantes semblera minime si l'on ne voit pas que le but de cette approche est finalement de permettre à l'utilisateur d'attribuer à chaque composante un nom qui soit en relation avec la nature des variables qui la déterminent et

ainsi d'être mieux outillé pour interpréter la différence des projections de deux individus sur l'une quelconque des composantes principales ou, plus généralement, les différences qui apparaissent entre les groupes d'individus mis en évidence par une représentation dans l'espace des composantes principales.

DEUXIEME EXEMPLE :

Des données de Monsieur Grandjouan vont nous fournir un deuxième exemple.

Sur un groupe de 78 placettes forestières, on a noté la valeur de 14 variables qui donne la matrice de la page 37.

Les résultats du programme de classement sont les suivants :

Groupe 1	variables 4,5,6,1	variable rejetée 12
Groupe 2	variables 11, 12	variable rejetée 3
Groupe 3	variables 8,9,7,10	variable rejetée 14
Groupe 4	variables 2,3	variable rejetée 13
Groupe 5	variable 13,14	

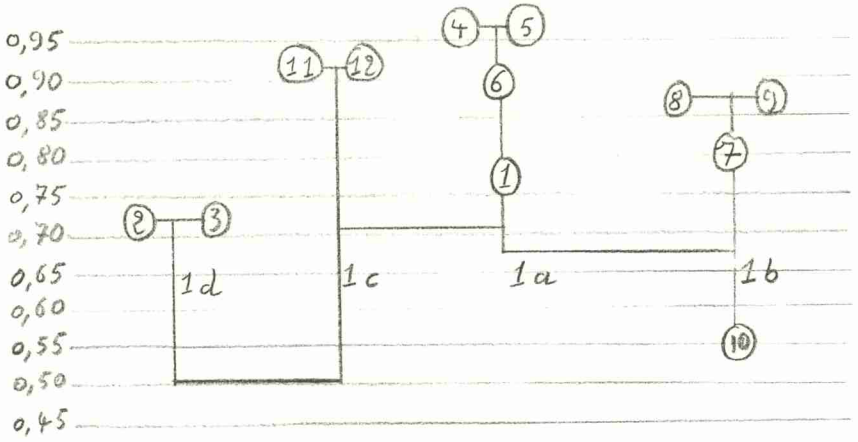
Nous remarquerons la sévérité du test qui fait quatre groupes des variables indicées de 1 à 12, alors que ces variables prises deux à deux donnent des coefficients de corrélations élevés dans leur ensemble. Nous sommes donc amenés à considérer, qu'en fait la matrice est constituée de deux groupes de variables, le premier groupe étant lui-même divisé en trois parties.

Groupe 1	1a	variables 4,5,6,1
	1b	variables 8,9,7,10
	1c	variables 11,12
	1d	variables 3,4
Groupe 2		variables 13, 14

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	
x_1	1													
x_2	0,2000	1												
x_3	0,5258	0,7367	1											
x_4	0,7685	-0,0958	0,3033	1										
x_5	0,7920	-0,0541	0,3323	0,9808	1									
x_6	0,7878	0,0906	0,4064	0,8931	0,9211	1								
x_7	0,4784	-0,1279	0,0957	0,6133	0,6514	0,6536	1							
x_8	0,4699	-0,2279	0,0995	0,7023	0,7183	0,7460	0,7903	1						
x_9	0,4485	-0,2001	0,0581	0,5833	0,6013	0,6137	0,6895	0,8485	1					
x_{10}	0,2305	-0,0328	0,1053	0,3379	0,3122	0,2841	0,5310	0,4435	0,5578	1				
x_{11}	0,6035	0,4207	0,5199	0,5637	0,6067	0,6629	0,2543	0,1393	0,0381	0,0926	1			
x_{12}	0,6467	0,3443	0,4679	0,6696	0,7069	0,7486	0,3510	0,3050	0,1900	0,1549	0,9316	1		
x_{13}	0,1824	0,2087	0,0263	0,0679	0,0286	0,1117	0,0568	0,0031	0,0855	0,2925	0,2052	0,1831	1	
x_{14}	0,3235	0,0977	0,0835	0,3494	0,2885	0,2511	0,2293	0,1678	0,2558	0,3309	0,1501	0,1939	0,6100	1

Fig. 1

Pour visualiser la configuration des variables du premier groupe, nous adopterons le procédé suivant : sur une échelle dont les barreaux sont des niveaux pour les valeurs des coefficients de corrélation, nous plaçons les deux variables qui présentent la plus forte corrélation. Nous cherchons ensuite le coefficient de corrélation immédiatement inférieur ; ou bien il concerne une variable déjà utilisée et nous introduisons la nouvelle au niveau voulu ; ou bien, il est relatif à deux nouvelles variables et nous commençons un nouveau groupe.



Sur ce shéma, nous remarquons que les groupes Ic et Id font pendant à Ib par rapport à Ia ce qui nous suggère, pour la représentation vectorielle des variables que les faisceaux de vecteurs représentants les variables Id et Ic d'une part, Ib d'autre part, sont placés de part et d'autre du faisceau Ia. Si, donc, nous admettons la présence d'une forte composante placée au centre du groupe - donc au voisinage du faisceau Ia - nous devons compter sur une deuxième composante qui opposera les groupes Ic et Ib.

Résultats de l'analyse des composantes principales effectuée sur la matrice construite avec les variables du groupe I

valeurs propres	6,390	2,586	0,651	1,109
pourcentage	53,25%	20,96 %	6,01 %	9,24 %
cumulé	53,25%	74,21 %	80,22 %	89,46 %
	Y_1	Y_2	Y_3	Y_4
x_4	0,365	0,050	- 0,042	- 0,220
x_5	0,375	0,031	- 0,065	- 0,221
x_6	0,378	- 0,021	- 0,113	- 0,122
x_1	0,329	- 0,132	- 0,235	- 0,051
x_8	0,304	0,318	- 0,243	0,091
x_9	0,268	0,347	- 0,192	0,252
x_7	0,292	0,256	0,106	0,166
x_{10}	0,177	0,223	0,636	0,552
x_{11}	0,262	- 0,378	0,393	- 0,193
x_{12}	0,301	- 0,297	0,353	- 0,205
x_2	0,042	- 0,492	- 0,116	0,484
x_3	0,178	- 0,410	- 0,348	0,414

La première composante apparaît bien comme la composante de groupe. Les variables 3,10 et 2 surtout tranchent par la faiblesse de leur participation tandis que le faisceau Ia est confirmé dans sa position centrale.

La deuxième composante oppose les effets des variables de Ib à ceux des variables de Ic et Id, tandis que le groupe Ia n'intervient presque pas.

La troisième variable oppose les variables 7,10,11,12 à l'en-

semble des autres, et la quatrième Ib et Id d'une part à Ia et Ic d'autre part.

Il faudrait maintenant donner à ces différentes composantes le nom qui correspond au phénomène biologique qu'elles mettent en évidence mais ce n'est pas mon propos.

Donnons enfin les résultats obtenus à partir des 14 variables On constatera les faibles variations des composantes déjà définies et l'apparition d'une composante nettement attachée au groupe II.

Valeurs propres	6,541	2,602	0,633	1,035	1,652
pourcentage	46,7%	18,6%	4,5%	7,4%	11,8%
cumulé	46,7%	65,3%	69,8%	77,2%	89,0%
	Y_1	Y_2	Y_3	Y_4	Y_5
x_4	0,360	0,057	- 0,135	0,203	- 0,105
x_5	0,368	0,040	- 0,098	0,164	- 0,152
x_6	0,371	- 0,015	- 0,050	0,071	- 0,126
x_1	0,327	- 0,129	- 0,320	0,051	- 0,030
x_8	0,296	0,322	- 0,080	- 0,171	- 0,083
x_9	0,265	0,345	- 0,121	- 0,262	- 0,062
x_7	0,287	0,257	0,211	- 0,184	0,003
x_{10}	0,182	0,208	0,523	- 0,377	0,353
x_{11}	0,259	- 0,375	0,397	0,216	- 0,080
x_{12}	0,296	- 0,294	0,359	0,219	- 0,091
x_2	0,046	- 0,496	- 0,059	- 0,460	0,143
x_3	0,175	- 0,403	- 0,269	- 0,495	- 0,043
x_{13}	0,080	- 0,100	0,096	0,212	0,660
x_{14}	0,151	- 0,004	- 0,396	0,229	0,585

Remarque : le changement de signe de Y_4 n'affecte pas l'interprétation qu'on peut en faire.

Ces considérations amènent quelques conclusions :

1°) Par une permutation judicieuse des variables, nous avons pu mettre en évidence, à partir de la matrice des corrélations, des groupes de variables qui nous renseignent aussitôt sur l'existence de composantes que nous avons appelées de groupe. Le cas limite où les variables ne formeraient qu'un seul groupe, qui donnerait une seule composante "générale" n'est évidemment pas exclu.

Nous avons vu au passage que, si la participation d'une composante à la variation totale est fonction de la valeur des coefficients de corrélation qui lient deux à deux les variables qui la définissent, elle est aussi fonction du nombre de ces variables, si bien que par un choix orienté des variables utilisées, il est possible de modifier l'importance relative d'une composante de groupe par rapport à une autre.

Au contraire, et c'est un problème qu'a abordé différemment Hotelling [17], si l'on suppose que l'expérimentateur dispose d'une population infinie de variables, dans laquelle il prend au hasard les variables sur lesquelles il va travailler, les groupes de variables obtenus sont une image des groupes de variables qui constitent la population. Si p est la probabilité de tirer une variable d'un groupe donné, on sait que la fréquence relative de l'apparition d'une variable de ce groupe est une variable aléatoire de moyenne p et d'écart type $\sqrt{\frac{p(1-p)}{n}}$, ce qui établit que lorsque n augmente la participation au total de la composante tend vers une constante comme $\frac{1}{\sqrt{n}}$ tend vers 0.

2°) Des procédés graphiques nous ont permis de mettre en évidence des

sous-groupes de variables et de placer à côté des composantes de groupe, un certain nombre de composantes secondaires. Ici encore, la part prise par une composante secondaire dans la variation totale est fonction du nombre de variables qui la déterminent si bien que l'on peut tirer les mêmes conclusions que pour les composantes de groupe.

3°) La méthode que nous avons proposée pourra, dans des cas très simples, permettre de donner les résultats d'une analyse des composantes principales sans qu'il soit besoin d'utiliser de gros moyens de calculs ; mais ce seront des cas d'exception. Plus généralement, elle doit conduire à mieux suivre la genèse des différentes composantes et donc, a être plus aptes à les interpréter puisque les ayant mieux comprises.

CHAPITRE QUATRIEME

Les différents calculs ont été effectués sur l'ordinateur IBM 1620 du laboratoire de Calcul de la Faculté. Je tiens à remercier vivement Monsieur Bernard Filliatre qui m'a initié à sa manipulation.

IV_1- Dans nos premiers travaux, nous avons calculé les valeurs propres et les vecteurs propres par la méthode proposée par Hotelling [17].

Il est connu ([26] p.25) que la rapidité avec laquelle cette méthode est susceptible de donner la $k^{ième}$ valeur propre lorsque, les valeurs propres étant classées dans l'ordre décroissant, les (k-1) premières ont été obtenues, est directement fonction de la rapidité avec laquelle les puissances successives du rapport $\frac{\lambda_k}{\lambda_{k-1}}$ tendent vers zéro.

Cette condition rend très difficile la convergence de la méthode dans bien des cas et nous a amené à choisir un procédé plus puissant, connu sous le nom de méthode de Jacobi.

Soit A, la matrice (symétrique dans notre cas) dont on veut calculer les valeurs propres. Supposons donnée une matrice T. la matrice $B = TAT^{-1}$ a les mêmes valeurs propres que A.
Le procédé va consister à choisir des matrices T_{ij} qui annulent les éléments non diagonaux (a_{ij}) de la matrice de telle sorte qu'après un certain nombre d'opérations la matrice

$$D = T_{i_k j_k} \; T_{i_{k-1} j_{k-1}} \cdots T_{i_2 j_2} \; T_{i_1 j_1} \; A \; T^{-1}_{i_1 j_1} \; T^{-1}_{i_2 j_2} \cdots \cdots T^{-1}_{i_{k-1} j_{k-1}} \; T^{-1}_{i_k j_k}$$

soit diagonale. Les valeurs propres sont alors les éléments de la diagonale et la matrice des vecteurs propres de D étant la matrice identité, la matrice des vecteurs propres de A est :

$$T^{-1}_{i_1 j_1} \; T^{-1}_{i_2 j_2} \cdots T^{-1}_{i_{k-1} j_{k-1}} \; T^{-1}_{i_k j_k} \; I$$

On définit T_{ij} par

$$t_{ii} = t_{jj} = \cos\theta$$
$$t_{ij} = -\sin\theta$$
$$t_{ji} = \sin\theta$$

Les éléments diagonaux autres que t_{ii} et t_{jj} sont égaux à 1 ;

Les éléments non diagonaux autres que t_{ij} et t_{ji} sont nuls.

Il est alors immédiat que $T_{ij}^{-1} = T_{ij}^{T}$. D'autre part si l'on marque d'une astérisque les éléments de $T_{ij}AT_{ij}^{T}$, on voit, par utilisation des lois sur la multiplication des matrices que :

pour $k \neq i$, $k \neq j$, $\ell \neq i$, $\ell \neq j$:

$$a^{*}_{k\ell} = a_{k\ell}$$

$$a^{*}_{ij} = a_{ij}(\cos^2\theta - \sin^2\theta) + \cos\theta \times \sin\theta(a_{ii} - a_{jj})$$
$$a^{*}_{ii} = a_{ii}\cos^2\theta + a_{jj}\sin^2\theta - 2a_{ij}\cos\theta\sin\theta$$
$$a^{*}_{jj} = a_{ii}\sin^2\theta + a_{jj}\cos^2\theta + 2a_{ij}\cos\theta\sin\theta$$
$$a^{*}_{kj} = a^{*}_{jk} = a_{kj}\cos\theta + a_{ik}\sin\theta$$
$$a^{*}_{ik} = a^{*}_{ki} = a_{ik}\cos\theta - a_{jk}\sin\theta$$

On devrait choisir θ de telle sorte que a^{*}_{ij} soit nul ce qui donnerait :

$$\frac{a_{ij}}{a_{ii} - a_{jj}} = \frac{\sin\theta\cos\theta}{\sin^2\theta - \cos^2\theta}$$

Soit $\dfrac{2a_{ij}}{a_{ii} - a_{jj}} = \operatorname{tg}(2\theta)$

Nous préfèrerons un procédé approché qui consiste à définir pour $a_{ii} \neq a_{jj}$:

$$g = \operatorname{tg}\left(\frac{\theta}{2}\right) = \frac{a_{ij}}{2(a_{ii} - a_{jj})}$$

Si $g < tg \frac{\Pi}{8} = 0,4142$, nous garderons cette valeur de g.

Si $g \geq tg \frac{\Pi}{8}$, nous prendrons $g = 0,41421$ $(\theta = \frac{\Pi}{4})$

Si $a_{ii} = a_{jj}$, nous prendrons aussi $g = 0,41421$

Nous définirons ensuite cos θ et sin θ par les expressions rationnelles

$$\sin\theta = \frac{2g}{1+g^2} \qquad \text{et} \quad \cos\theta = \frac{1-g^2}{1+g^2}$$

Avec cette valeur de θ, a^{*}_{ij} n'est pas exactement nul. On fera donc plusieurs itérations : ayant mis en mémoire une valeur ε considérée comme pratiquement nulle, on travaillera à chaque itération sur le plus grand des a_{ij} jusqu'à ce que tous soient inférieurs à ε.

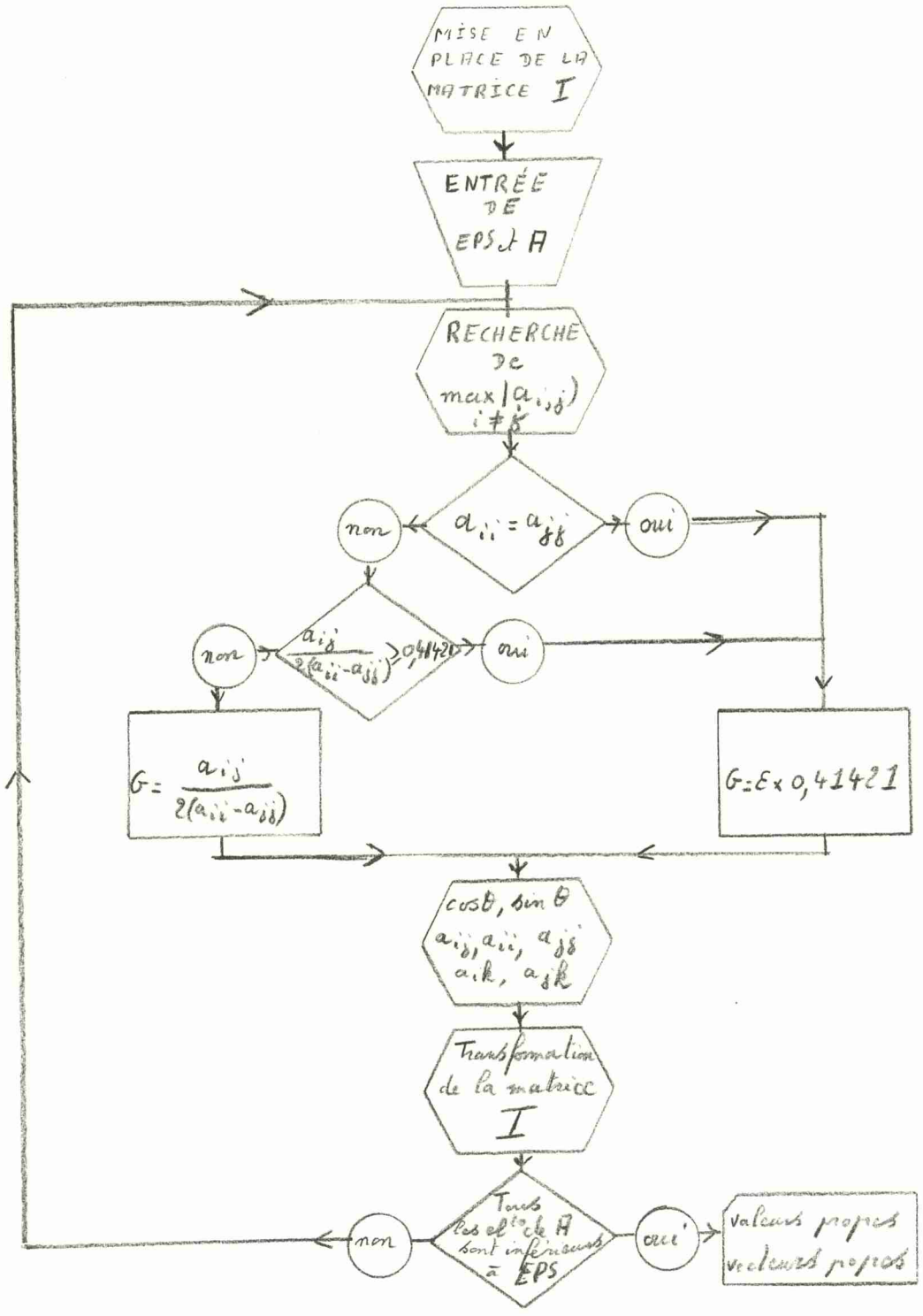
Notons que le choix des matrices T_{ij} permet d'obtenir les vecteurs propres comme les lignes de la matrice :

$$T_{i_k j_k} \; T_{i_{k-1} j_{k-1}} \cdots T_{i_2 j_2} \; T_{i_1 j_1} \; I$$

On trouvera ci-joint l'organigramme et le programme Fortran I relatifs à cette méthode.

IV_2- L'organigramme et le programme d'organisation de la matrice des corrélations découlent directement des explications données en III_2.

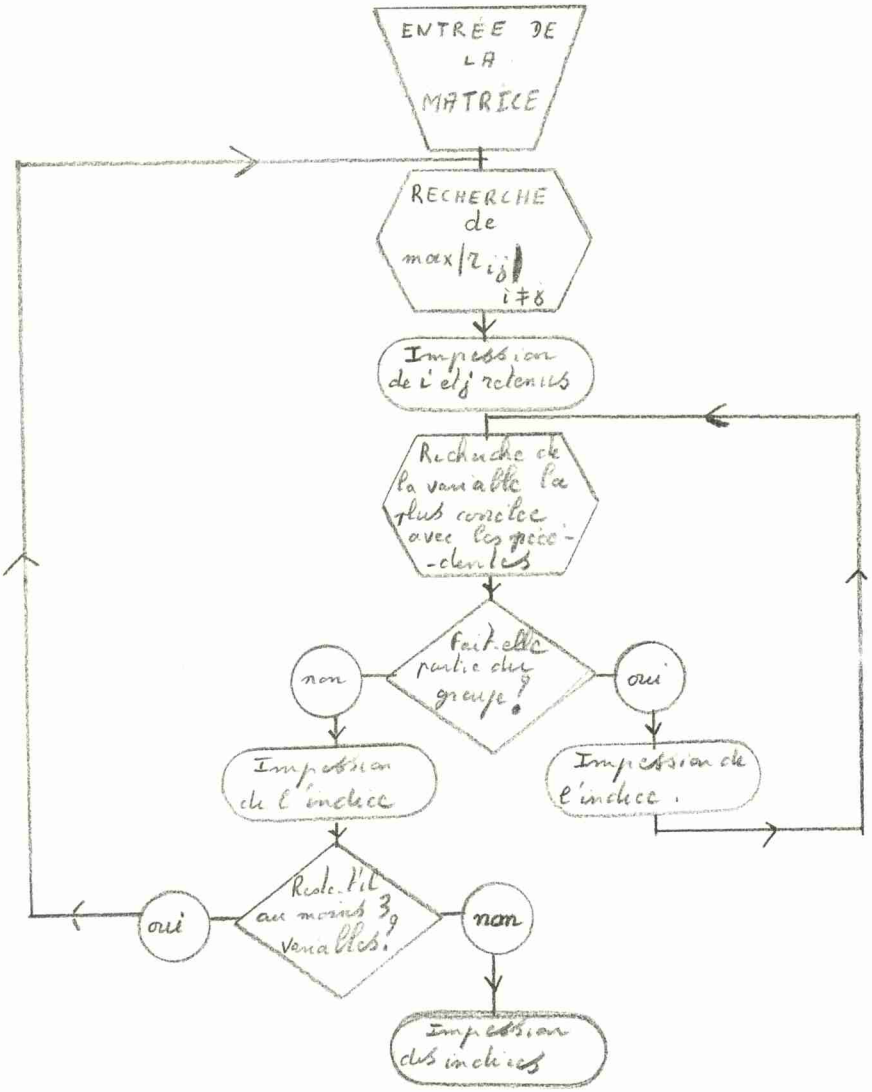
ORGANIGRAMME DE LA METHODE DE JACOBI



```
06600 C      METHODE DE JACOBI-VALEURS PROPRES-RANG MAXIMUM=21-
06600        DIMENSION A(233),V(21,21)
06600    100 READ1,EPS,N
06636        DO 60 I=1,N
06648        DO 60 J=1,N
06660        IF(I-J)61,62,61
06728     61 V(I,J)=0.
06812        GO TO 60
06820     62 V(I,J)=1.
06904     60 CONTINUE
06976        EPS=EPS/10.
07012        E=0.
07036        K2=N*(N+1)/2
07096        DO 10 K=1,K2
07108        READ2,I,J,AA
07156        IF(I-J)11,12,11
07224     11 IF(ABS(AA)-E)12,12,13
07292     13 E=ABS(AA)
07316     12 K1=J*(J-1)/2+I
07388     10 A(K1)=AA
07472     40 IK=1
07496        DO 27 J=2,N
07508        JJ=J-1
07544        DO 28 I=1,JJ
07556        K=J*(J-1)/2+I
07628        IF(ABS(A(K))-E)28,28,21
07720     21 IK=2
07744        K1=I*(I-1)/2+I
07816        K2=J*(J-1)/2+J
07888        IF(A(K2)-A(K1))31,32,31
08004     31 G=A(K)/2./(A(K2)-A(K1))
08160        IF(ABS(G)-0.41421)33,34,34
08228     32 G=A(K)
08276     34 G=0.41421*G/ABS(G)
08348     33 CS=(1.-G*G)/(1.+G*G)
08468        SN=2.*G/(1.+G*G)
08564        AA=A(K)*(CS*CS-SN*SN)+CS*SN*(A(K1)-A(K2))
08828        AB=A(K1)*CS*CS+A(K2)*SN*SN-2.*A(K)*CS*SN
09092        A(K2)=A(K1)*SN*SN+A(K2)*CS*CS+2.*A(K)*CS*SN
09368        A(K1)=AB
09416        A(K)=AA
09464        DO 20 K=1,N
09476        IF(K-J)23,20,24
09544     23 K2=J*(J-1)/2+K
09616        IF(K-I)22,20,25
09684     22 K1=I*(I-1)/2+K
09756        GO TO 26
09764     24 K2=K*(K-1)/2+J
09836     25 K1=K*(K-1)/2+I
09908     26 AA=A(K2)*CS+A(K1)*SN
10040        A(K1)=A(K1)*CS-A(K2)*SN
10208        A(K2)=AA
10256     20 CONTINUE
```

```
T0292          DO 70 K=1,N
T0304          AA=V(J,K)*CS+V(I,K)*SN
T0508          V(I,K)=V(I,K)*CS-V(J,K)*SN
T0784       70 V(J,K)=AA
T0904       28 CONTINUE
T0940       27 CONTINUE
T0976          GO TO (30,40),IK
T1052       30 E=E/10.
T1088          IF(E-EPS)50,50,40
T1156       50 DO 51 I=1,N
T1168          J=I*(I-1)/2+I
T1240       51 PUNCH3,I,A(J)
T1336          DO 80 J=1,N
T1348          DO 80 I=1,N
T1360       80 PUNCH4,J,I,V(J,I)
T1540          GO TO 100
T1548        1 FORMAT(F10.0,I2)
T1576        2 FORMAT(I3,I3,F10.0)
T1608        3 FORMAT(2HL(,I3,2H)=,E15.8)
T1658        4 FORMAT(2HV(,I3,1H,,I3,2H)=,E15.8)
T1722          END
```

ORGANIGRAMME DU PROGRAMME "CLASSEMENTS"



```
06600 C      CLASSEMENTS
06600        DIMENSION A(21,21),L(21)
06600   100  READ1,N,NI
06636        FNI=NI
06672        KN=N*(N-1)/2
06732        E=0.
06756        NN=N-1
06792        NT=0
06816        DO 10 I=1,N
06828     10 A(I,I)=1.
06948        DO 20 K=1,KN
06960        READ2,I,J,A(I,J)
07068     20 A(J,I)=A(I,J)
07248   200  S=0.
07272        SS=0.
07296        DO 30 I=1,NN
07308        IF(A(I,I))31,30,31
07424     31 II=I+1
07460        DO 30 J=II,N
07472        IF(A(J,J))30,30,33
07588     33 IF(ABS(A(I,J))-E)30,30,32
07716     32 IM=I
07740        JM=J
07764        E=ABS(A(I,J))
07848     30 CONTINUE
07920        PRINT3,IM
07944        PRINT3,JM
07968        A(IM,IM)=0.
08052        A(JM,JM)=0.
08136        MT=NT+1
08172        L(MT)=IM
08220        MT=NT+2
08256        L(MT)=JM
08304        NL=2
08328   300  MT=NT+NL
08364        IF(MT-N+1)81,82,400
08444     82 NT=MT
08468        GO TO 73
08476     81 IM=NT+1
08512        E=0.
08536        DO 40 I=1,N
08548        IF(A(I,I))41,40,41
08664     41 DO 50 J=IM,MT
08676        K=L(J)
08724     50 S=S+ABS(A(I,K))
08856        IF(S-E)51,51,52
08924     52 E=S
08948        IE=I
08972     51 S=0.
08996     40 CONTINUE
```

```
09032         FK=0.
09056         Z=0.
09080         DO 60 J=IM,MT
09092         FK=FK+1.
09128         K=L(J)
09176      60 Z=Z+(LOG(1.+ABS(A(IE,K)))-LOG(1.-ABS(A(IE,K))))/2.
09488         Z=Z/FK
09524         Z=Z+1.96/SQRT(FNI-3.)
09596         A(IE,IE)=0.
09680         DO 90 I=1,N
09692         IF(A(I,I))91,90,91
09808      91 ZZ=(LOG(1.+ABS(A(IE,I)))-LOG(1.-ABS(A(IE,I))))/2.
10072         IF(Z-ZZ)62,90,90
10140      90 CONTINUE
10176      63 NL=NL+1
10212         MT=NT+NL
10248         L(MT)=IE
10296         PRINT3,L(MT)
10344         SS=0.
10368         GO TO 300
10376      62 NT=NT+NL
10412         A(IE,IE)=1.
10496         IF(NT-N+1)71,73,73
10576      71 PRINT4,IE
10600         E=0.
10624         GO TO 200
10632      73 MT=NT+1
10668         DO 80 I=1,N
10680         IF(A(I,I))87,80,87
10796      87 L(MT)=I
10844         PRINT3,L(MT)
10892         GO TO 400
10900      80 CONTINUE
10936     400 DO410 I=1,N
10948     410 A(I,I)=1.
11068         DO 420 I=1,N
11080         DO 420 J=I,N
11092         K1=L(I)
11140         K2=L(J)
11188     420 PUNCH5,I,J,A(K1,K2)
11368         GO TO 100
11376       1 FORMAT(I2,I3)
11404       2 FORMAT(I3,I3,F10.0)
11436       3 FORMAT(I3)
11458       4 FORMAT(10H++++++++++,15HVALEUR ESSAYEE ,I3)
11544       5 FORMAT(I3,I3,F10.6)
11576         END
```

TABLE DES REFERENCES

Traités généraux de statistique.

[1] Anderson T.W (1964) - An introduction to multivariate Statistical Analysis - John Wiley and Sons

[2] Cramer H. (1954) - Mathematical methods of statistics - Princeton university Press.

[3] Dugué D (1958) - Traité de Statistique théorique et appliquée - Masson et Cie.

[4] Keeping E.S (1962) - Introduction to statistical inference - Van Nostrand Company.

[5] Kendall M.G (1961) - A course in multivariate analysis - Griffin's Statistical monographs and courses.

[6] Radhakrishma Rao C (1952) - Advanced statistical methods in biometric research - John Wiley an Sons.

[7] Snedecor G.W (1956) - Statistical methods - The Iowa State University Press.

[8] Wilks S.S (1950) - Mathematical statistics - Princeton University Press.

Publications spécifiques

[9] Bartlett M.S (1951) - The effect of standardization on a X^2 approximation in factor analysis. Biométrika 38,337-344.

[10] Bartlett M.S (1954) - A note on the multiplying factors for various χ^2 APproximations - J. Roy.Stat.Soc. B 16,296-298.

[11] Benzecri J.P (1964) - Sur l'analyse factorielle des proximités- Publications de l'I.S.U.P. Vol XIII, 235-282.

[12] Burt C (1949). Alternative methods of factor analysis and their relations to Pearson's method of Principal Axes -Brit.J.Psychel. Statist, Seet 2, 98-121.

[13] Colloques internationaux du C.N.R.S. (1955) - l'Analyse factorielle et ses applications - C.N.R.S.

[14] Dagnélie P. (1960) - Bulletin du Service de la carte phytogeographique - Série B - carte des Groupements Végétaux - Tome V - Fasc. 1,7,69 et Fasc. 293, 190.

[15] DEBAZAC E.F et TOMASSONE R. (1965) - Contribution à une étude comparée des Pins méditerranéens de la Section Halepensis - Ann. Des Sciences Forestières. Tome XXII - Fasc.2.215-256.

[16] Harmann H.H (1960) - Modern Factor Analysis - The university of Chicago Press.

[17] Hotelling H (1933) - Analysis of a complex of statistical variables into principal components - J. Educ. Psychol. 24; 417-441;498-520.

[18] Lawley D.N (1956) - Tests of significance for the latent roots of covariance and correlation matrices - Biométrika 43;128-136.

[19] Lawley D.N et Maxwell A.E (1963) - Factor Analysis as a statistical method - Butterwords.

[20] Pearson K (1901) - On lines and planes of Closet fit to Systems of points in space - Phil - Mag 2 ; 6ième séries ; 557-572

[21] Tomassone R (1965) - L'analyse des composantes principales - Note Scientifique n°1 - C.N.R.F.- Station de Biométrie.

[21 bis] Rao C.R (1964) - The use and interpretation of principal component analysis in applied Research- Sankhya - Séries A- Vol.26-Part 4.

Problèmes divers

[22] Bass (1956) - cours de Mathématiques Masson et Cie

[23] Durand E (1961) - Solutions numériques des équations algébriques T.II - Masson et Cie.

[24] Fox L (1964) - An introduction to numerical linear Algebra. Clarendon Press - Oxford.

[25] Lentin A. et Rivaud J (1961) - Eléments d'algèbre moderne - Vuibert.

[26] Notes on Applied Science n°16 (1961) - Modern computing Methods - Her Majesty's stationery office - London.

[27] Wilkinson J.H (1963) - Rounding Errors in Algebraic Processes - Her Majesty's stationery office - London.